Use of displays with packaged statistical programs

by *W. J. DIXON*
University of California
Los Angeles, California

During the past few years there has been a great increase in the use of packaged statistical programs. These programs are prepared in a general form. For example, a regression program will allow the user to specify:

the number of variables being introduced as a data set
the number of cases
the choice from this data set of the dependent variable
the choice of some subset of the input variables to be the independent variables
the type of input (cards vs. tapes, etc.)
the transformation of any variable in the data set
the construction of new variables for some stated function of the input variables
the stepwise computation of regression function
the priority with which variables may be considered for introduction into the regression equation
the plotting of residuals vs. various input variables.

The output provided includes means, standard deviation, correlation, and at each step of the regression process regression coefficients, partial correlation, and an analysis of variance table for the regression.

At the end of the process a list of residuals and the plots mentioned above are provided.

A regression program of this generality will enable a user to compute a wide variety of problems, as well as the standard regression problem, e.g., analysis of variance, contrasts either orthogonal or dependent, analysis of covariance, discrimination, etc.

A library of statistical programs such as the BMD series developed at UCLA gives the user a selection of several different regression packages as well as a variety of data screening, analysis of variance, multivariate analysis and other statistical procedures, all providing a considerable degree of flexibility to cover a wide variety of problems.

In biomedical applications, as well as in other fields, studies with large amounts of data are analyzed sequentially. Various analyses are carried out to choose among various possible dependent or outcome variables and to relate these variables to the many independent variables. Analyses are made to screen the data for accuracy, including search for outliers, etc. The data are examined for linearity, homogeneous covariance matrices, etc. The investigator can be assisted by the computer during this phase, but he must still participate actively in directing the process.

If these programs are used at a computing facility providing batch processing and providing these programs in its internal program library the successive steps in the analysis can be accomplished by submitting a call for a specific program along with a specification of the program parameters and a data deck or tape. Each stage may be submitted separately and the statistician will usually adjust later steps in the analysis from the outcome of earlier computations. The same program, or several different programs, may be used. It may be possible also to store the data in the computer and call it from the file whenever the next analysis is desired. The return of each portion of the analysis may be a matter of hours, perhaps in some cases, a matter of minutes. The data analyst may designate different dependent variables or modify the choice or priorities of the dependent variables, transform the variables, include or exclude cases, or more generally he may choose the definitions of strata.

This analytical search is in some ways analogous to the stepwise regression itself. The art of data analysis has not yet matured to the point where the investigator can specify in advance the algorithms which might be brought to play in this analytical process much less to specify their sequence.

The development of interactive computer systems will greatly enhance the researcher's capacity to ana-

lyze his data. In such systems summary information can be supplied to the user at a console and he can supply commands to the computer from his console. The investigator may therefore intervene in the strategy of analysis, indicating that a run with the same program should be repeated with a specification of new or modified parameters, etc. Thus, this type of system reduces the turnaround time for successive stages of the analysis. Some investigators have found earlier interactive systems to be too restrictive in their available modes of input and output and would prefer a capability additional to that of the typewriter console to provide card input and fast printer output. However, it should be noted that output which requires a great amount of time to read may as well be provided off-line. Also, a very large number of cards may be more easily handled at a central computer.

Computing systems to handle interactive consoles, as well as providing adequate capacity for executing large statistical programs, will soon be generally available. So far the statistician has been required to choose between interactive capacity and capacity to do a comprehensive analysis. He, of course, requires both capacities concurrently. He also needs aid in visualizing the interrelationships of variables and the goodness-of-fit of various statistical models.

Statistical graphs

Many experts in data analysis have always used graphical methods to aid their analysis of data. One often hears directives of these experts to their assistants something like, "go thou and *plot* your data." The plots and charts frequently do not survive the process of report writing and publication, but have played an important part in the analytical process itself.

These charts have employed various colors or symbols to identify strata and lines have been drawn, dotted, dashed and perhaps wiggly. Question marks annotate extreme values. Regression lines or frequency contours may be drawn in alternate forms with and without suspect cases or with and without doubtful subgroups. Various forms of graph paper are used and curves are plotted with standard error regions.

The data analyst who has taken the computer unto himself may have dropped some of these time consuming graphical procedures in favor of various computer aided analyses of alternate approaches. He still, however, seeks graphical output from the programs he uses and still very likely employs some graphical methods in plotting his next analytical move.

Recently the interactive computing systems have been introducing television type screens to supply more rapid output at the user's console. These scopes may also be equipped with a light pen to be used for communicating signals from the user to the computer. These signals may be used as commands to steer the course of the analysis by supplying parameters ,changes in parameters, choice of subroutines and identification of data points, etc. The light pen is slightly larger than a fountain pen and affects the designation of information on the screen by sensing the screen regeneration at a particular position.

The users console may have available also a typewriter, instruction keys and perhaps card input, or more luxuriously such additional input media as paper tape, magnetic tape, disc packs and film. When the console is connected with a computing system that has a well developed operating system, program library and file service, the user may build effective programs by virtue of having his commands executed within a reasonable length of time, having available many packaged service and analytical programs as he works at the console. A few computing facilities are approaching this capacity at the present time. These facilities provide much of what is needed for effective data analysis.

A major difference will be the need to operate at a computer language level at least one step higher than that required for the programmer. The statistician will not in general be able to make sufficiently rapid progress with his analysis if he must do any significant amount of programming while he is "on-line."

Simple scopes which provide only character displays can satisfy needs of the programmer but will not provide the graphical needs of the statistician. Scopes are available which provide plotting capabilities and vector or line drawing features. When a scope of this type is coupled with a light pen to provide a simple form of identification and communication concerning the results observed on the scope and a function keyboard to initiate commands to subroutines and to supply parameter values, etc., the hardware capacity is available for the statistician to construct graphs and to interact with an analysis in progress in the computer in a "plotting" or graphical mode.

The next step in providing an operative system for him is to design the supporting software. The overall system will need to be serviced at several levels of programming. The following paragraphs indicate some of the statistician's needs.

Data storage

A basic data file will be assumed to be structured to an extent that one may arrange the data into a *data matrix* whose columns represent variables (i.e., types of measurements) and whose rows represent individuals or cases on whom the measurements have been made.

It is not assumed that every measurement is available for every case, but that various types of "missingness" are either coded in the original data or provided by a computer program.

Operations on the data file

One or more variables may be selected by name or number. Operations available on this variable (or on several variables) include functional transformation by a function supplied at the console. These new variables may be added to the data matrix or, if necessary, replace other variables in the data matrix to conserve storage space. Stratification of cases can be accomplished by forming submatrices based on categorizing statements involving the data entries for one or more variables. The above features provide, for example, the capacity to compute residuals from a regression function based on the same or other observations. Submatrices or extensions to matrices may be stored as derived files.

It is frequently desirable to perform a transformation on one or more variables where the functional form is known but for which not all parameters are either known or computable from the given data. One may wish to examine the effects of the transformation in a regression or discrimination problem and note the effect on the prime analysis of various selections of the parameters of the transformation. This calls for a convenient way of providing a sequence of values to one or more parameters and to watch at the scope the effect on the plotted outcome of the computations. In some cases it may be sufficient to display conversion indicators in numerical form and guide the process accordingly.

Parameter pacing

A simple replacement of the parameter values one by one at the console is frequently not sufficient to the task. A pacing subroutine needs a starting value and an increment to provide a succession of inputs so that the progression of events may appear as rapidly at the scope as they may be provided by the main computer. This subroutine also provides an "increase" and "decrease" instruction to accelerate or decelerate pacing and a backspace instruction to avoid restarting the parameter search when an optimum point has been passed.

Scope plots

The statistician uses plotting on paper in many ingenious ways to increase his understanding of various characteristics of the data. A change of scale, e.g., from arithmetic paper to logarithmic paper or the use of probit or logit paper for cumulative proportions can be accomplished on the computer by transformations on the data.

The ability to plot data pairs (x, y) on paper including the use of dots or other characters to represent strata is directly transferable to the scope. Although the use of colors for strata is not generally available on today's computers, the use of motion provides an even richer visual presentation. If a particular stratification merely represents an ordered categorization based on a third variable the identification of strata may be viewed as the capacity to introduce a third dimension to the two cartesian coordinate variables for the plane face of the scope.

Let us examine for a moment what can be displayed on the scope. The simplest representation of one variable can be dots along a line or, in two dimensions, a point in the plane. A second variable can be introduced in the plane by locating a line segment at x-horizontally and using a line segment for y (sometimes called a histogram). A third dimension may be added to a point (x, y) by changing the dot size or displaying a line segment whose center is (x, y). The value of a third variable can be indicated by the length of the line. A third variable may also be introduced by using a short line segment of fixed length whose center is positioned at (x, y) but whose angle of inclination represents a third variable.

Stereo representations can be provided which give true stereo with the use of viewing lenses and the impression of stereo by configurations changing shape as though rotating in space. When regeneration is spaced in time and applied successively to subgroups of the points (or symbols, characters, or line segments) attention is drawn successively to each subgroup in turn. The subgroup designation may be a third variable. Subgroups may be differentially highlighted by the frequency or duration of regeneration. If sufficient system support can be given to the console scope, line segments may be given either a metronome motion or rotation whose frequency or speed represents an additional variable. Lengths of line segments may be automatically scaled down from maximum length to give a length proportional to the third variable. The slope of a line segment may represent a residual scaled by the maximum residual or represent a third variable scaled to cover its range between $+1$ and -1. These facilities can be supplied with relative ease to the graphical system. The pictorial display may be alternated with or overlayed by the usual numerical output of the statistical program. The "printed" and pictorial presentation can also be located at different places on the scope.

Background grid

A grid of the usual graph paper type can be easily provided but since color is not usually available, the

grid lines should be of lesser intensity than the data points. A more effective mode for the user is provision of regeneration to the background only at specified times or on call. The user must be able to specify line widths and different frequencies of regeneration. For finer work, provision for second or fifth lines preferably at a different intensity or width should be provided.

Before giving further details on the specification of other modes for displaying additional dimensions, we give examples of the types of variables which one may wish to display.

Since each observation of multivariate data on continuous variables is often conceptually visualized as representing a point in higher dimensional space, the capacity to represent more dimension in the two-dimensional frame is desirable. Categorical data which is ordered can be treated similarly. Special attention may be needed for unordered categories. Preliminary analyses may provide additional dimension to the plotted data. If the plotted points are means, the additional dimension of sample size and standard deviation are important. If percentages are computed on categorical data, the sample size and proportion non-responding are important, etc.

Case identification

The *light* pen may be used to identify a point representing a case. This can cause the computation to be repeated removing this case, showing the resulting changes in the display. A simple example is the display of two alternate regression lines computed with and without one or more designated points (cases). Upon identification of a point and specification of a particular variable, the user can request a histogram showing the entire distribution of that variable with a flashing or highlighted location of the identified case in the distribution.

Program control by light pen

Sequencing of the analysis can be made by light pen or by function keys. Experience has shown that the light pen is easier to use because of the appearance of memory cues and choice alterations on the scope. The pacing of parameters can also be guided by light pen from a choice of the pacing parameters themselves.

Selection of strata

The specification of various strata for analysis can be readily accomplished by calling a selection subroutine whose descriptors are prepared for the particular study. Consider for example a tumor registry which has been prepared using various codes for age, sex, diagnosis, treatment, etc. This subroutine can exhibit first these prime variables. The touch of the light pen to the word "age" appearing on the screen will cause the specific code for age to appear, perhaps showing that age is coded to 5 year age groups. The light pen can be used to touch the desired age groups and return to the original list and proceed. In some cases (depending on which variable is chosen) a choice tree of several levels may be specified by a succession of choices determined by light pen. The computer program constructs a Boolean selection statement which can be used to operate on the data file.

It is fairly obvious that the various forms of case or strata selection can be combined with the various forms of variable selection, the various forms of graphic presentation and the various statistical models of analysis to provide a very powerful tool.

The literature holds fairly complete documentation of the analytical methods available. Systems of data analysis programs are described by W. J. Dixon as Chapter 3 in *Computers in Biomedical Research* edited by Ralph W. Stacy and Bruce Waxman, Academic Press, New York, 1965.

Examples of packaged graphics programs developed thus far using some of the above features are:

1) Simple plot and regression. This program provides data matrix operations including selection and transformation, provides scatter plots with case removal or addition with adjusted regression.

2) Stepwise regression with control of selection and exclusion of variables, transformation, etc.

3) Spectral analysis with series selection, transformation, filter construction and providing spectrum, cospectrum, phase relations, etc.

4) Non-linear regression with specification of function parameters and boundary conditions and control of iterations.

5) File search with code assisted Boolean specification for constructing subfiles with control of descriptions of subgroups.